# Couplage serré
# Multiprocesseur symétrique
# (SMP - Symmetric MultiProcessor)

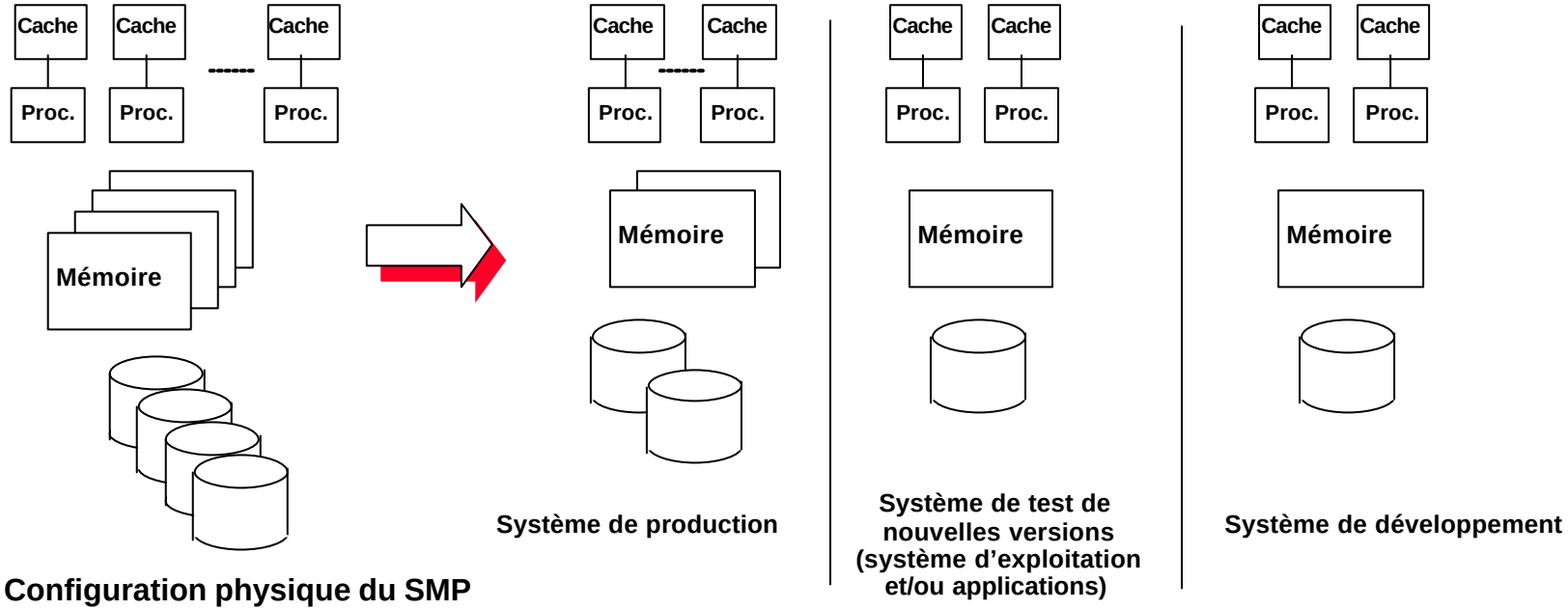
# Exemples

# Caractéristiques « Haut de Gamme »

- **La plupart des systèmes à grand nombre de processeurs présentent des caractéristiques communes telles que celles rencontrées traditionnellement sur les systèmes propriétaires :**
  - **Partitionnement des systèmes :**
    - **Statique**
    - **Dynamique**
  - **Redondance de certains composants (ventilateurs, alimentation électrique, processeurs)**
  - **Livraison de systèmes avec le nombre maximal de processeurs et « libération » des processeurs à la demande**
  - **Facturation fondée sur les processeurs réellement utilisés**
  - **Échange de composants à chaud (« Hot Plug »)**
  - **….**

# Partitionnement des SMP

- **Concept hérité des mainframes (répartition des ressources entre plusieurs systèmes indépendants)**




Configuration physique du SMP

Système de production

Système de test de nouvelles versions (système d'exploitation et/ou applications)

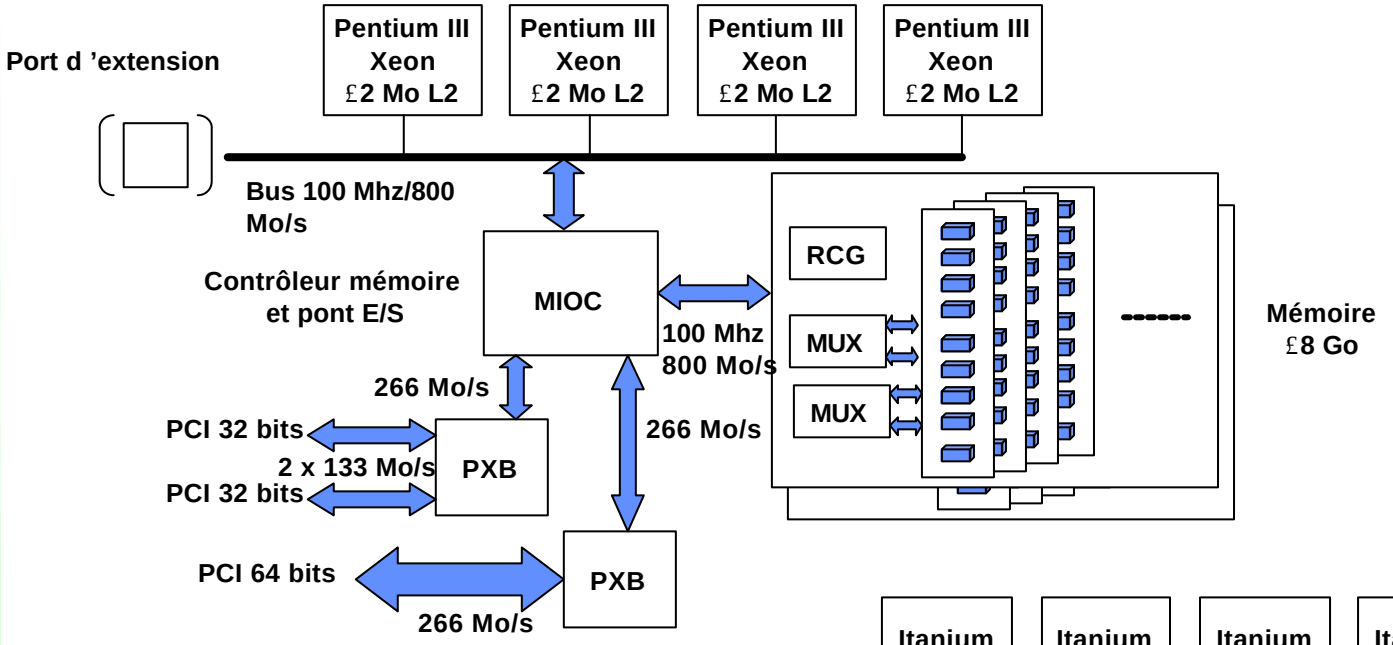
Système de développement

# Architecture flexible SMP/Cluster

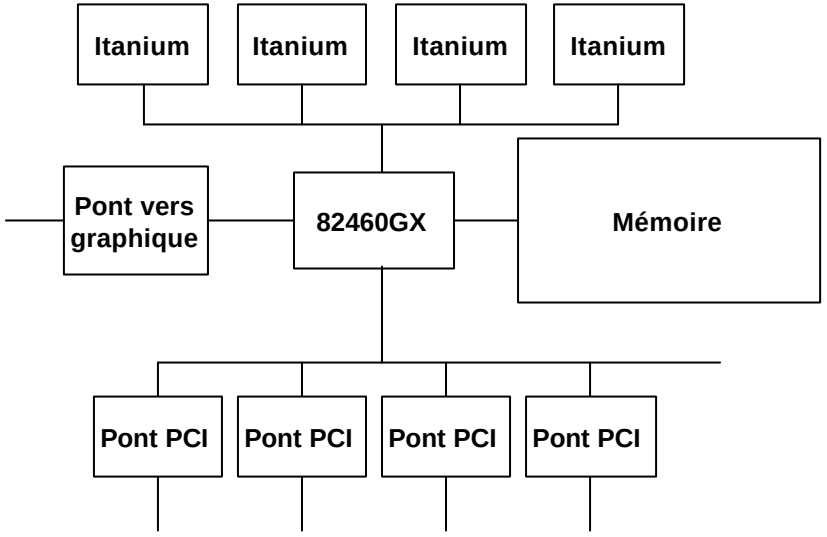
- **Objectif :**
  - **Adapter la configuration du système au besoin des applications et non pas les applications à l'architecture du système**
  - **Approche : concept d'un SMP à grand nombre de processeurs et mécanismes de partitionnement du SMP en plusieurs SMP indépendants**
  - **Bénéfice : une seule architecture couvre une plage étendue de besoin au prix d'une conception un peu plus complexe**
  - **Exemple de cette approche : CMP (Cellular MultiProcessing) d'Unisys**

# SMP à nombre modéré de processeurs (<= 8 processeurs)

# SMP à nombre modéré de processeurs




***Architecture SHV Pentium III (Source Intel)***

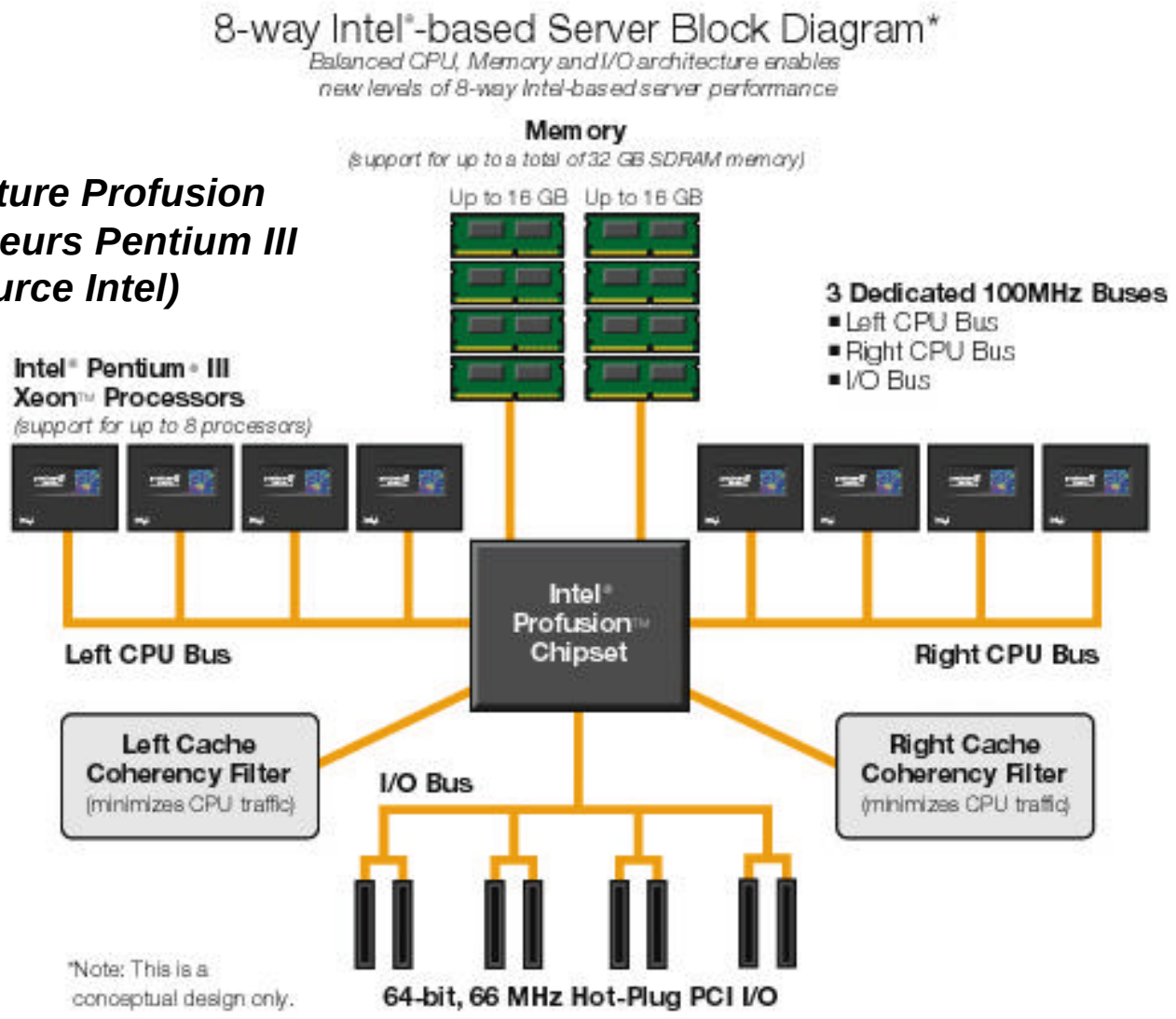



***Architecture SHV Itanium (Source Intel)***

***SHV = Standard High Volume***

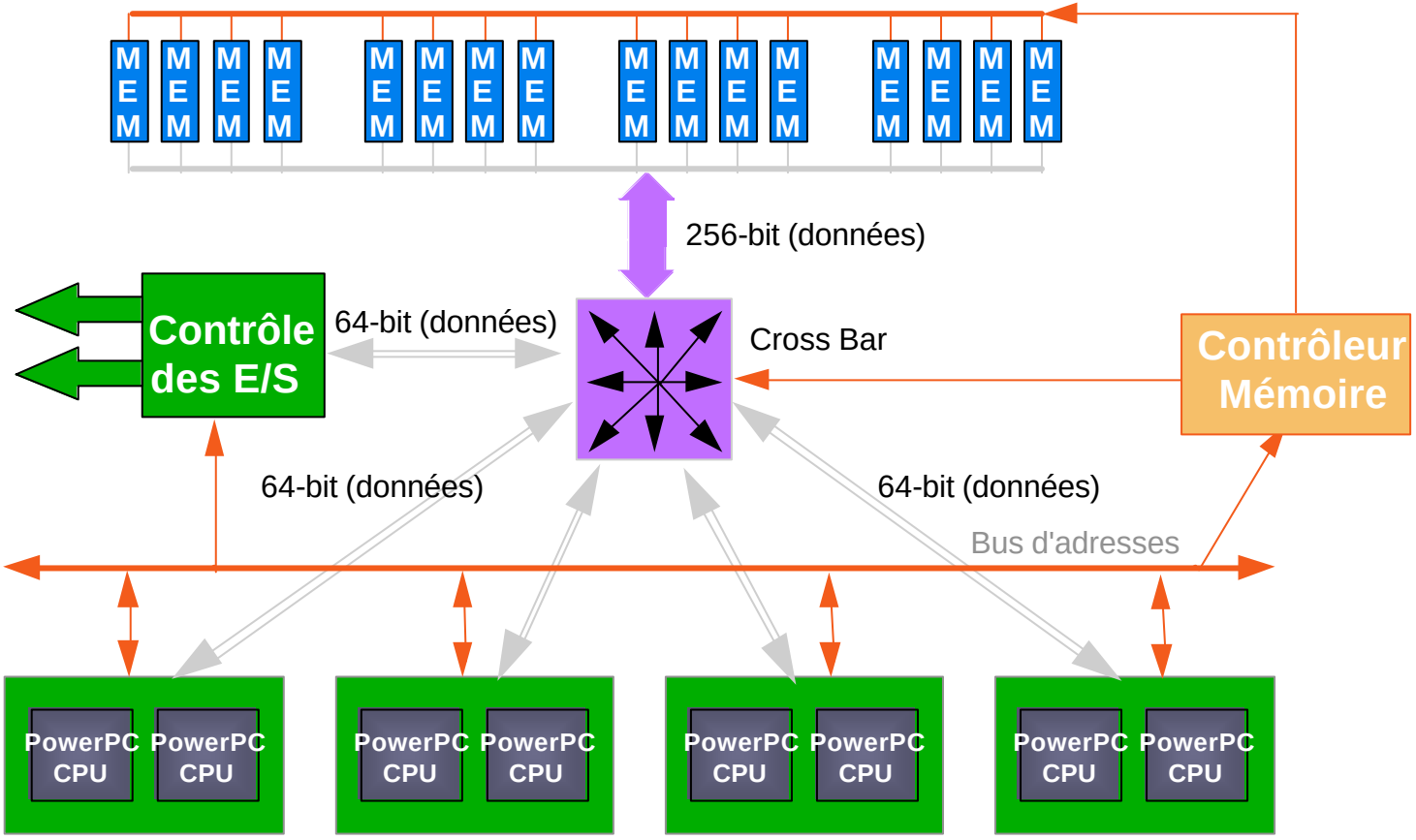
# Extension de l'architecture SHV

***Architecture Profusion***
***8 processeurs Pentium III***
***(Source Intel)***



# Architecture de type cross-bar

## *Architecture PowerScale, système Escala 8 processeurs Power PC (Source Bull)*



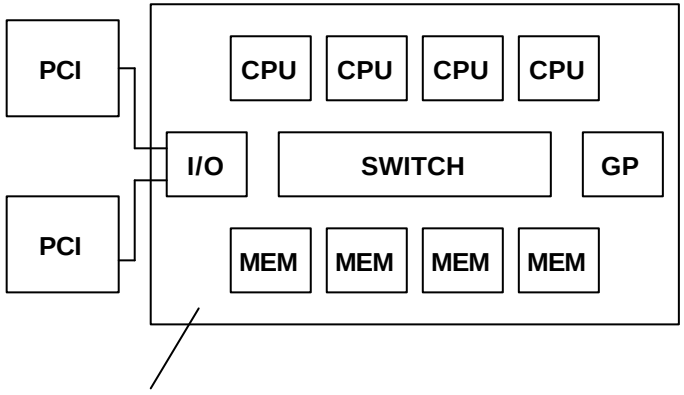
# SMP à grand nombre de processeurs (>= 8 processeurs)

# SMP à grand nombre de processeurs

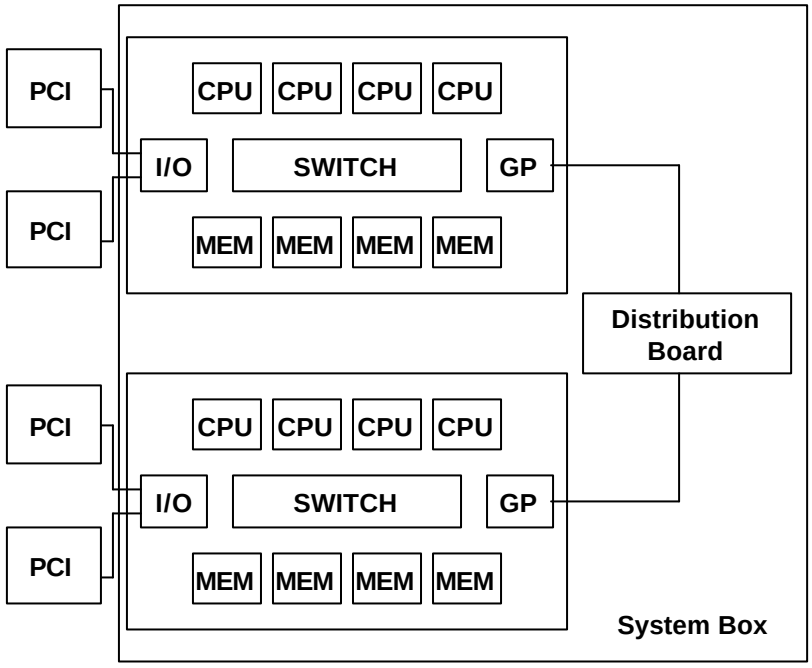
- **Nécessité de répartir la mémoire et les processeurs en plusieurs sous-ensembles**
- **Deux grandes familles en fonction du rapport des temps d'accès aux mémoires (locales et distantes)**
  - **UMA (Uniform Memory Access) ou proche de UMA (Nearly UMA) si le rapport est inférieur à 2**
  - **NUMA (Non Uniform Memory Access) si le rapport est supérieur à 2. Avec le respect de la cohérence de cache, les architectures à temps d'accès non uniforme sont appelées CC-NUMA (Cache Coherent - Non Uniform Memory Access)**

***Note : Il s'agit ici d'une règle empirique. On considère qu'avec un rapport inférieur à 2, il n'est pas obligatoire que le système d'exploitation soit notablement modifié pour tenir compte de la localité*.**

# Compaq AlphaServer GS80



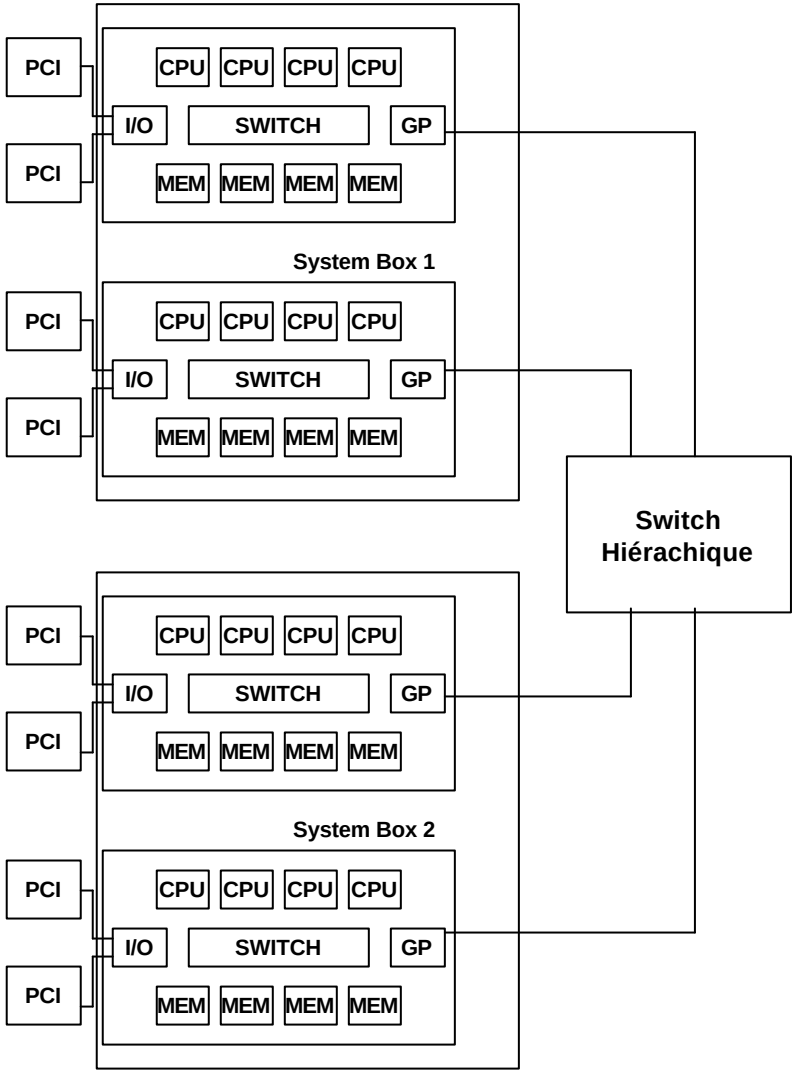

QBB Quad Building Board 32 Go




GS80 : 8 processeurs, 64 Go, 12.8 Go/s, 56 slots PCI/3.2 Go/s

# Grands SMP - Compaq(2)

## Compaq AlphaServer GS160/GS320




**GS160 :**

- **16 processeurs, 128 Go, 25.6 Go/s, 112 slots PCI/6.4 Go/s**

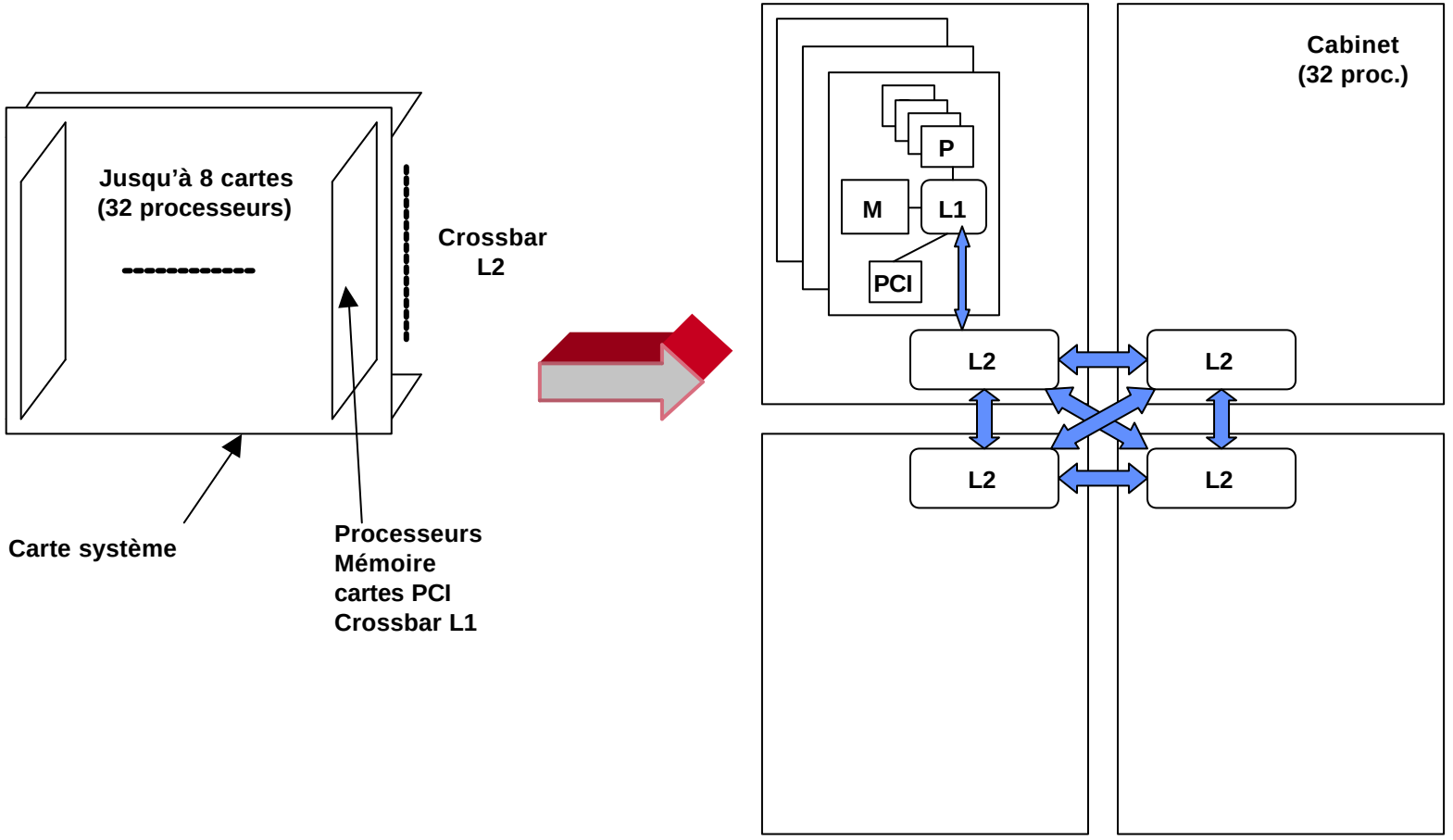
**GS320 :**

- **32 processeurs, 256 Go, 51.2 Go/s, 224 slots PCI/12.8 Go/s**

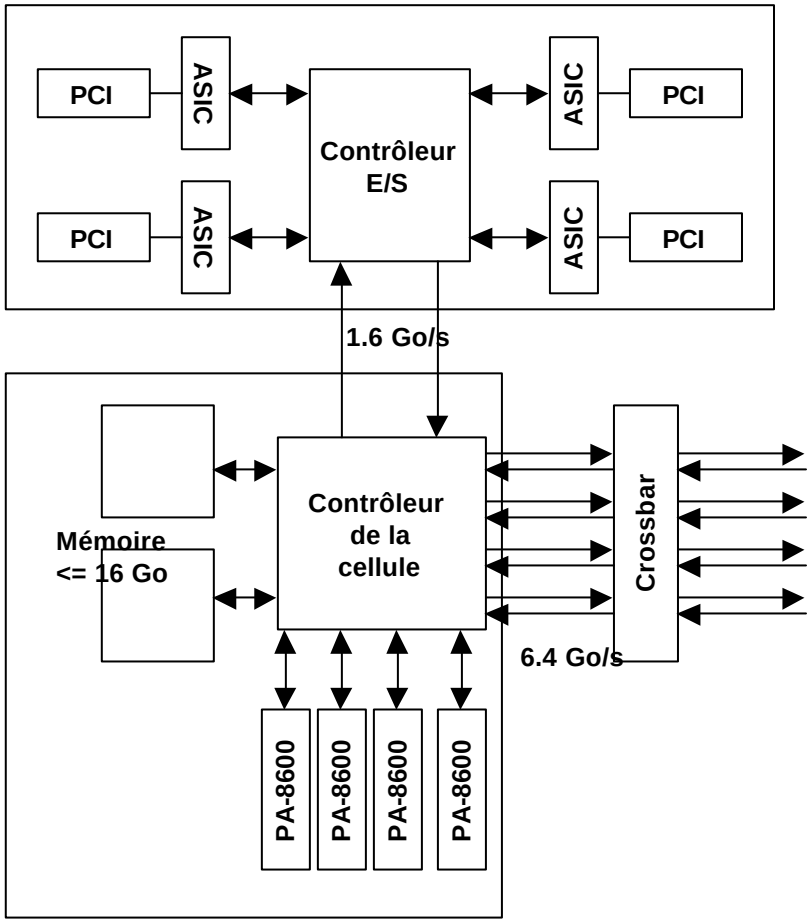
# Grands SMP - Fujitsu

- **PRIMEPOWER 800/1000/2000**
  - **Systèmes allant jusqu'à :**
    - **16 (modèle 800), 32 (modèle 1000) ou 128 (modèle 2000) processeurs SPARC64 GP (450 Mhz, 8 Mo cache L2)**
    - **512 Go de mémoire**
    - **192 cartes PCI**
  - **Architecture de type hiérarchie de crossbars :**
    - **L1 interne à la carte quadri-processeur**
    - **L2 entre cartes et entre modules :**
      - **3 parties : adresses, état des granules de cache, données**
      - **8000 signaux au maximum (90 cm max, 225 Mhz)**
      - **Latence mémoire : 300 ns à travers le crossbar de niveau 2 (!)**
    - **Cohérence fondée sur le snooping**
      - **Débit maximal de snooping = 57.6 Go/s**
      - **Débit maximal de données = 81.9 Go/s**
  - **Remplacement à chaud de différents éléments :**
    - **Processeurs et Mémoires**
    - **Cartes PCI**
    - **Unités d'alimentation électrique**
    - **Ventilateurs**
    - **Cartes de contrôle système**
    - **Disques**
  - **Partitionnement dynamique (15 maximum)**

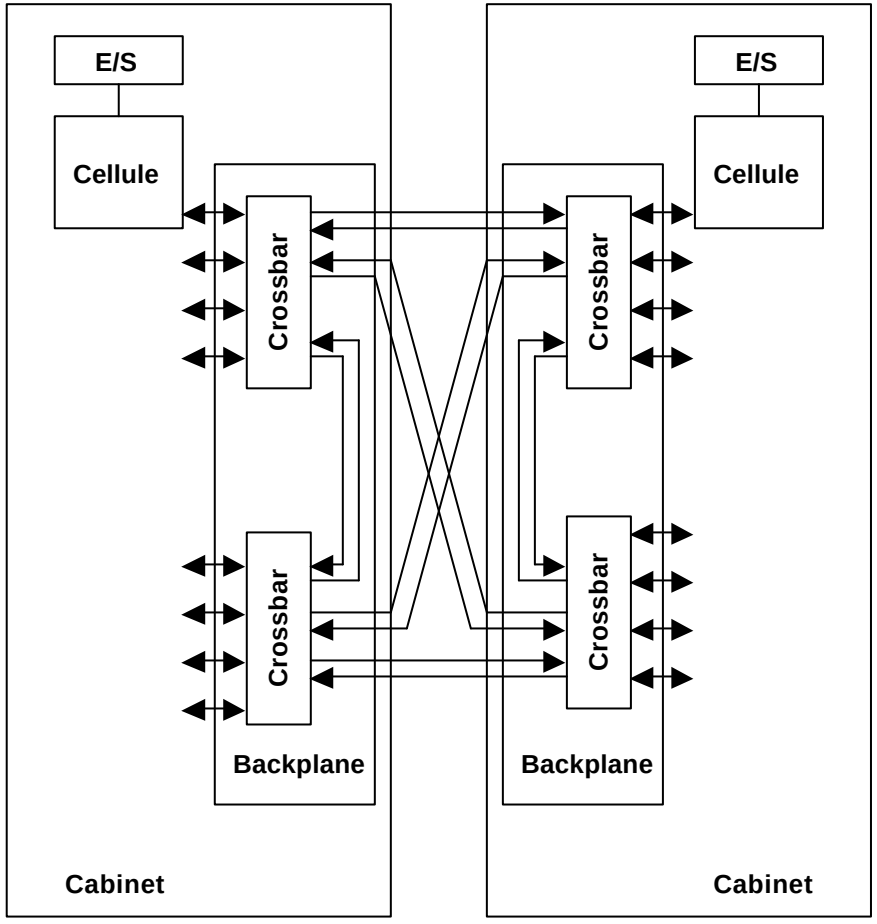
## Illustration : PRIMEPOWER 2000



# HP Superdome : Architecture de la cellule



# HP Superdome : jusqu'à 64 processeurs



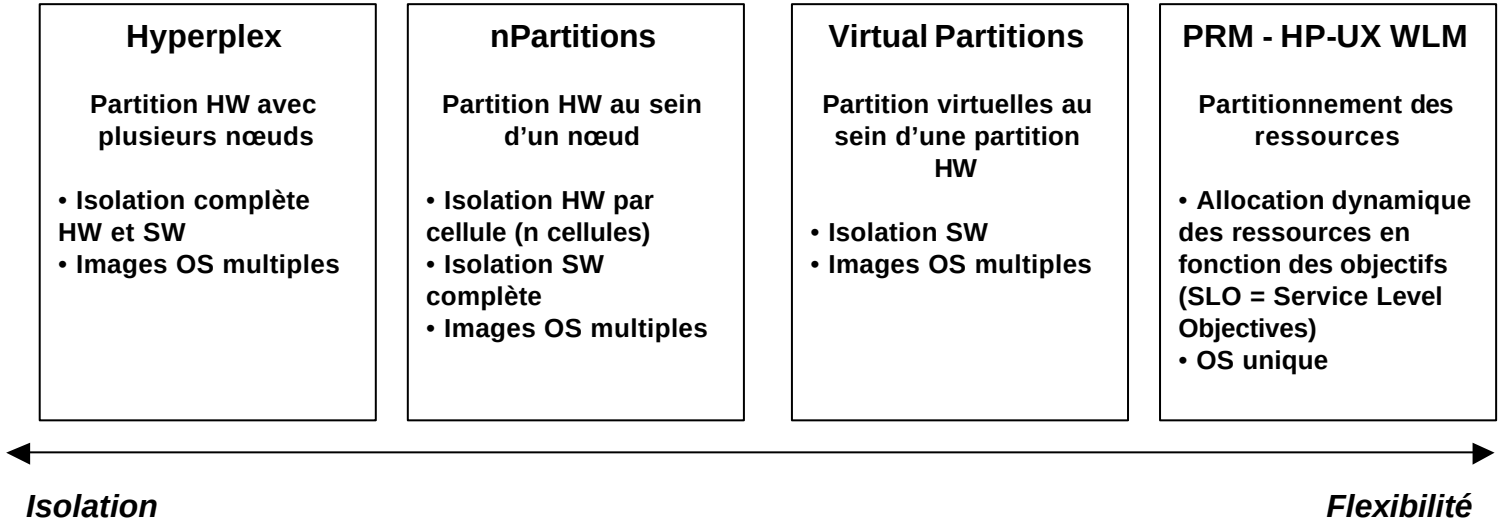

**Débit Crossbar :**

- **16 processeurs = 12.8 Go/s**
- **32 processeurs = 25.6 Go/s**
- **64 processeurs = 51.2 Go/s**

**Latence mémoire moyenne en charge :**

- **au sein de la cellule : 260 ns**
- **sur le même cossbar : 320 à 350 ns**
- **au sein d'un même cabinet : 395 ns**
- **entre cabinets : 415 ns**

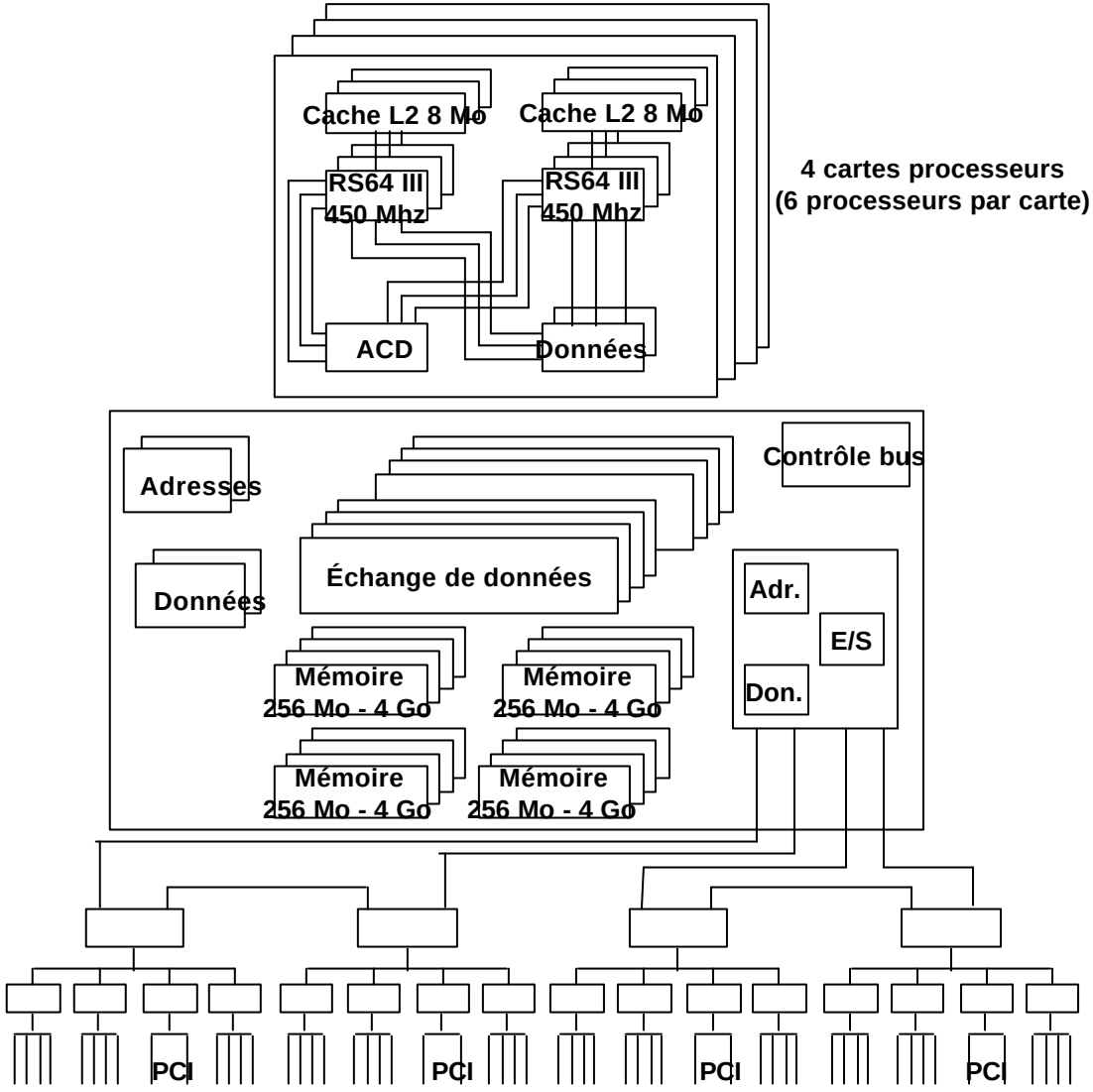
## HP Superdome : Partitionnement




## Haute disponibilité

- N+1 et Hot Swap pour alimentation, ventilateurs, cartes processeur et PCI

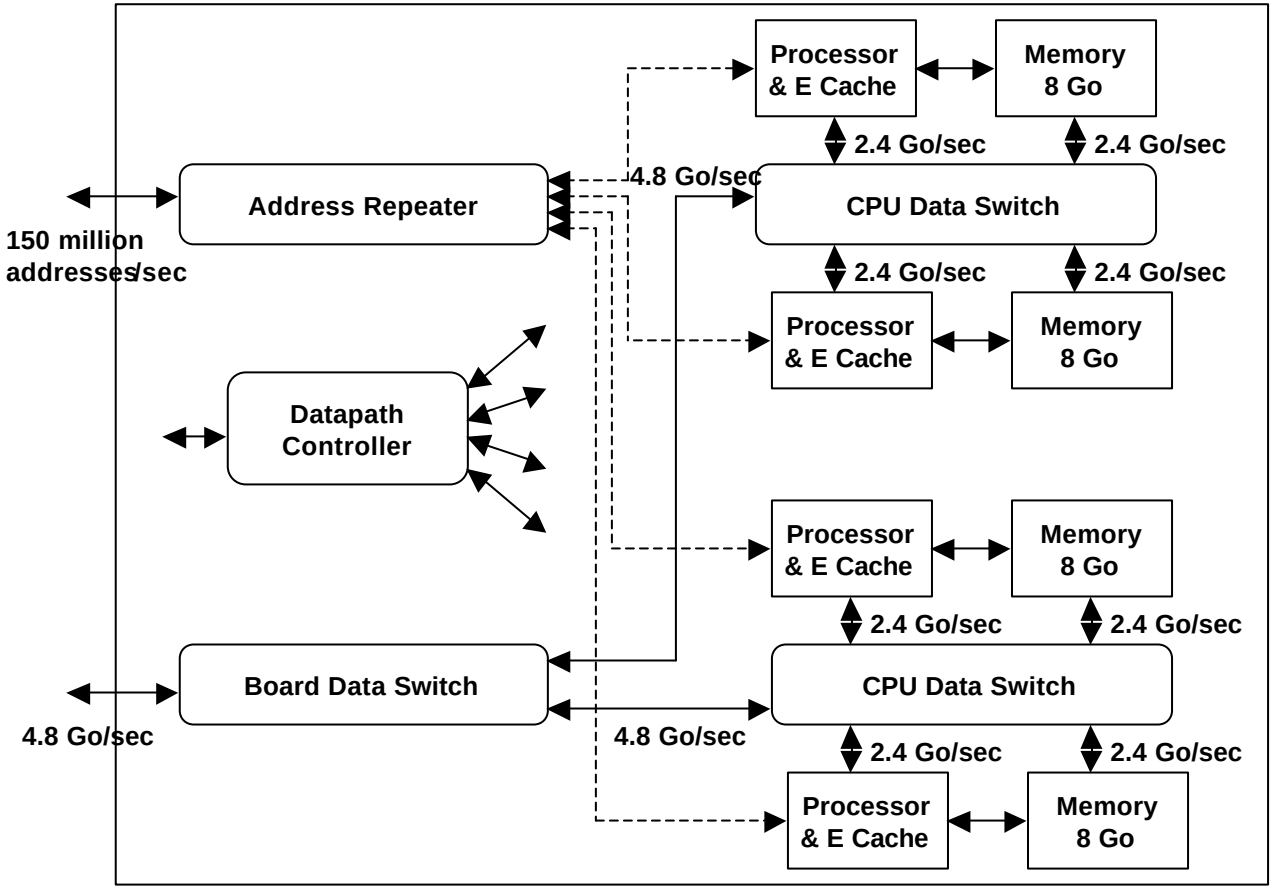
# IBM RS/6000 Enterprise Server Model S80



# Grands SMP - IBM AIX(2)

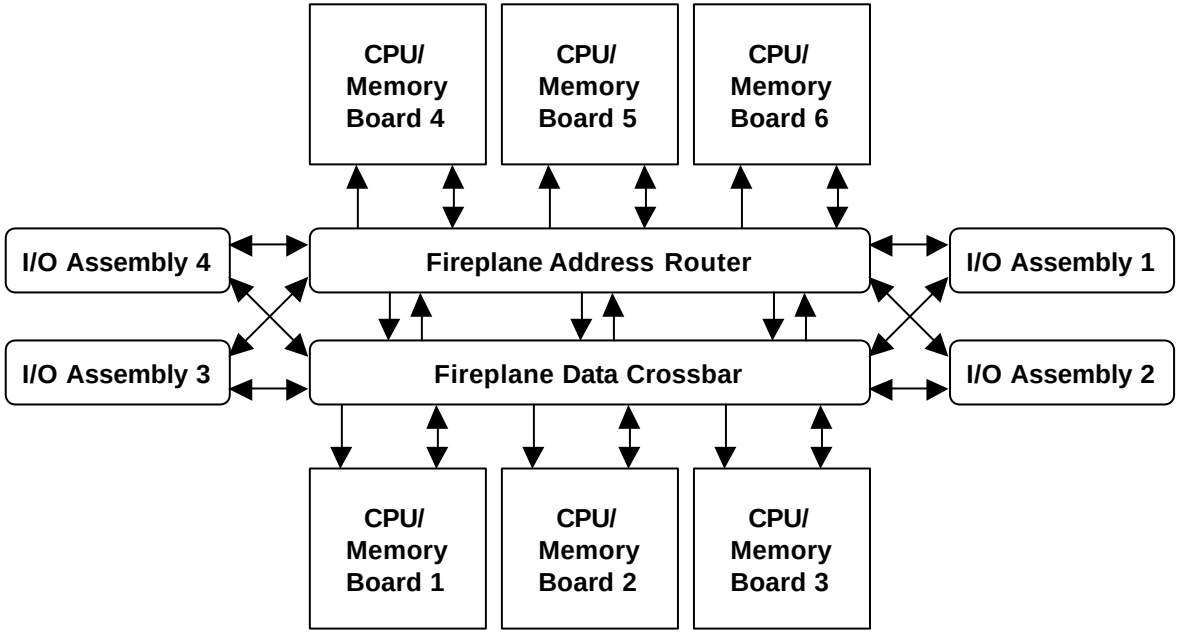
- **Haute disponibilité**
  - **Redondance de type N+1 et échange à chaud pour les éléments d'alimentation électrique et les ventilateurs**
  - **Disques Hot Plug**
  - **Partitionnement :**
    - **???**

- **Sun Fire 3800-4800-6800 Servers**
  - **Architecture de la carte de base**
    - **Fondée sur UltraSPARC III**
    - **<= 8 Mo de cache L2 (E cache)**



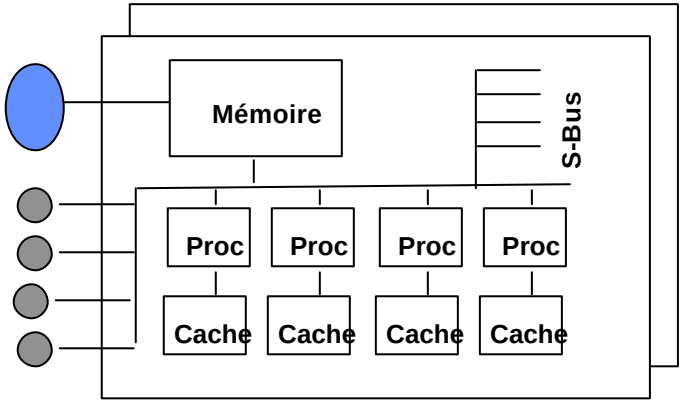
# Grands SMP - Sun(2)

- **Architecture du Sun Fire 6800 (24 processeurs)**
  - **Cohérence fondée sur le snooping (diffusion des adresses)**
  - **Temps d 'accès à la mémoire :**
    - **180 ns local (carte)**
    - **240 ns sur une autre carte**
  - **Débit de données = 67.2 Go/s**
  - **Hot plug :**
    - **Cartes Processeur/Mémoire, Alimentation électrique, Ventilateurs, Disques, I/O, Cartes PCI, …**
  - **Partitionnement dynamique du système (4 domaines sur le 6800)**



# Sun Ultra Enterprise 10000

**Gigaplane-XB**
**<12.8 G octets/s**
**Liaison intercartes**
**XB-interconnect**
**(cross-bar données)**
**4 bus d'adresse**




**Jusqu'à 16 cartes (64 processeurs)**

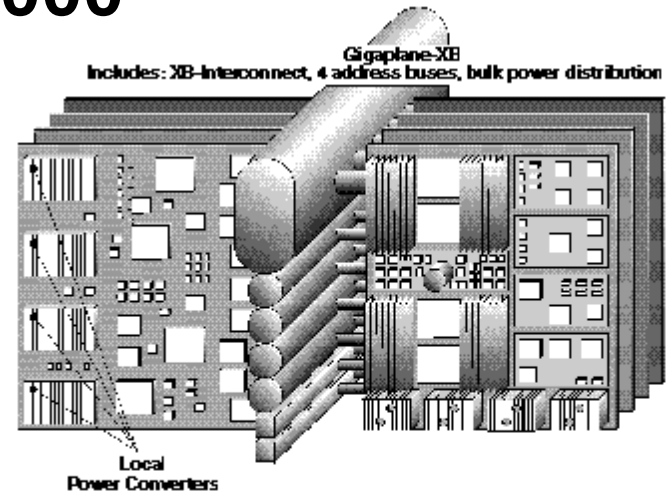



*Figure 2-2 System Boards with the Gigaplane-XB Interconnect*

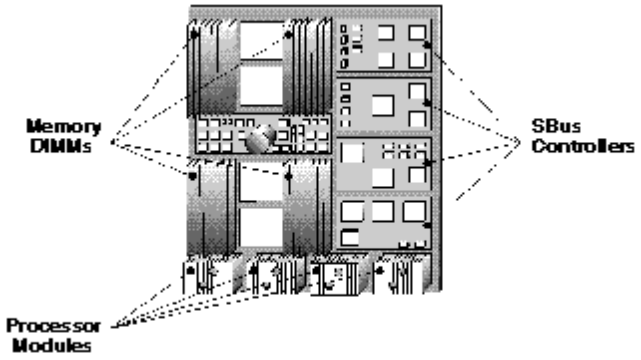



*Figure 2-1 System Board*

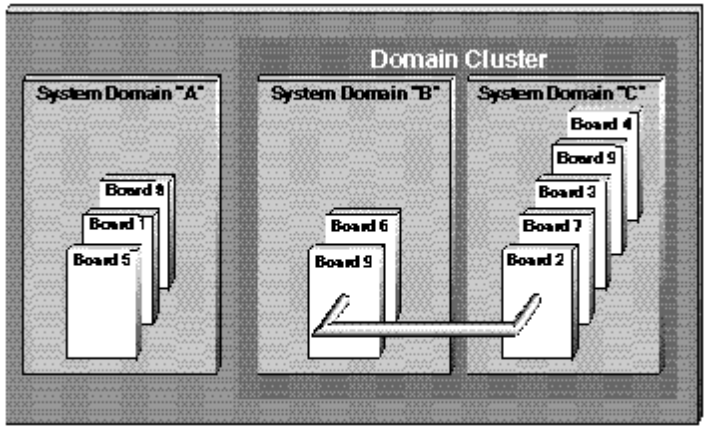
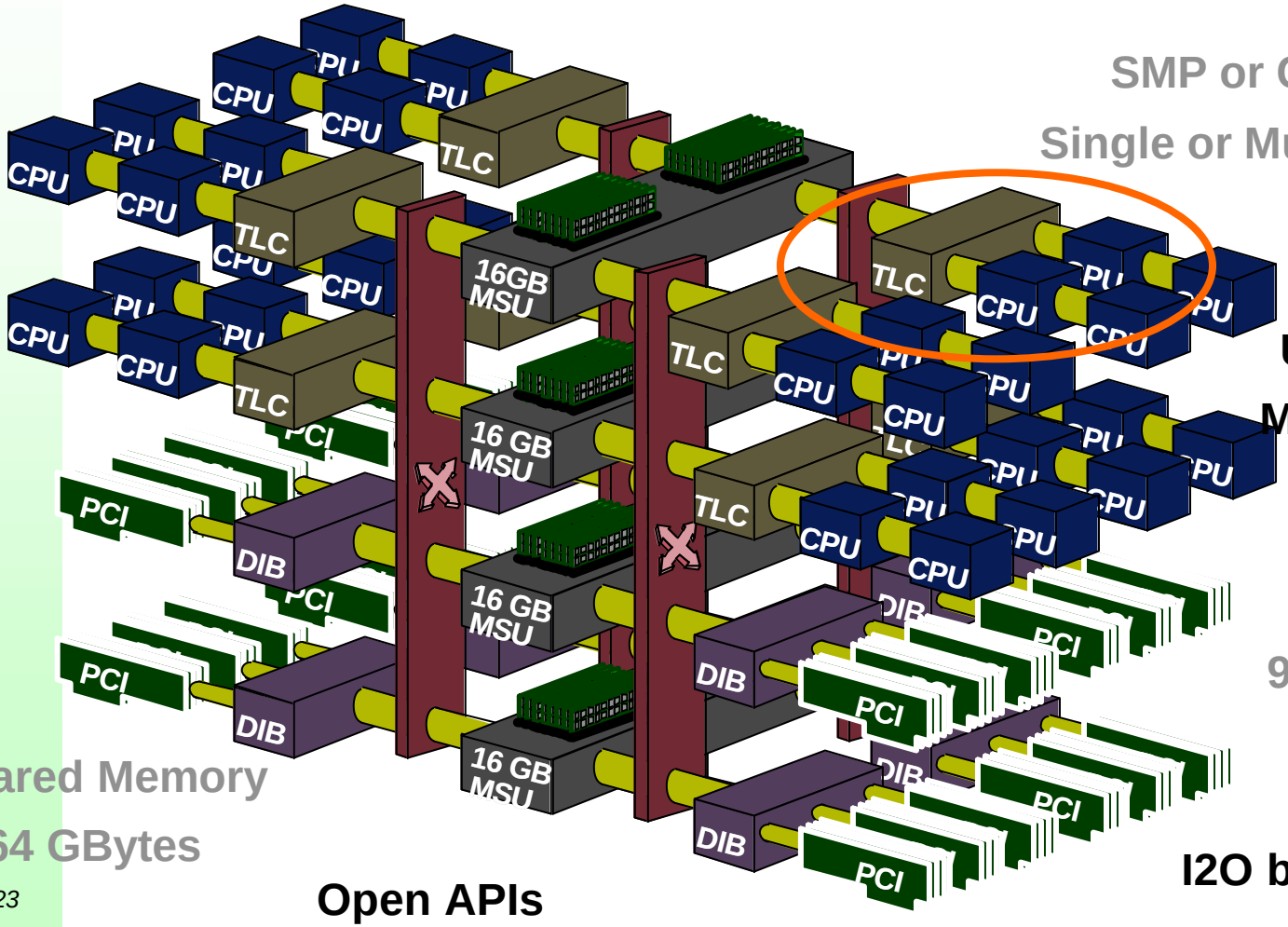



*Figure 2-3 Dynamic System Domains*

**Source Sun**

***Concept de domaine :***
***Partitionnement du système "à la mainframe"***

# Unisys CMP - Technologie de la plate-forme




SMP or Clusters

Single or Multiple OS's

IA32, IA64

Unisys CMOS

MCP or OS2200

96 PCI slots

I2O boards

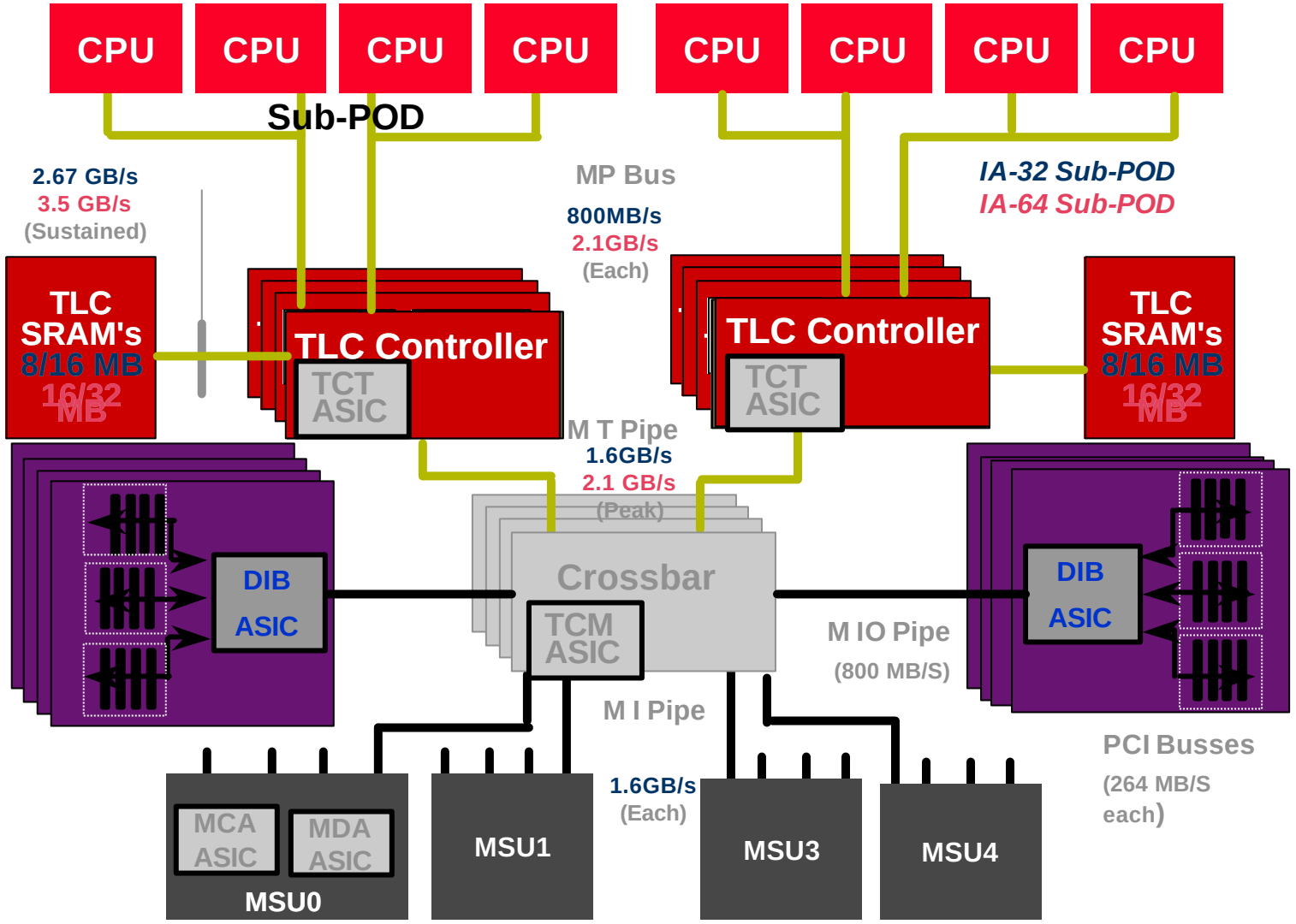
Shared Memory

<=64 GBytes

Open APIs

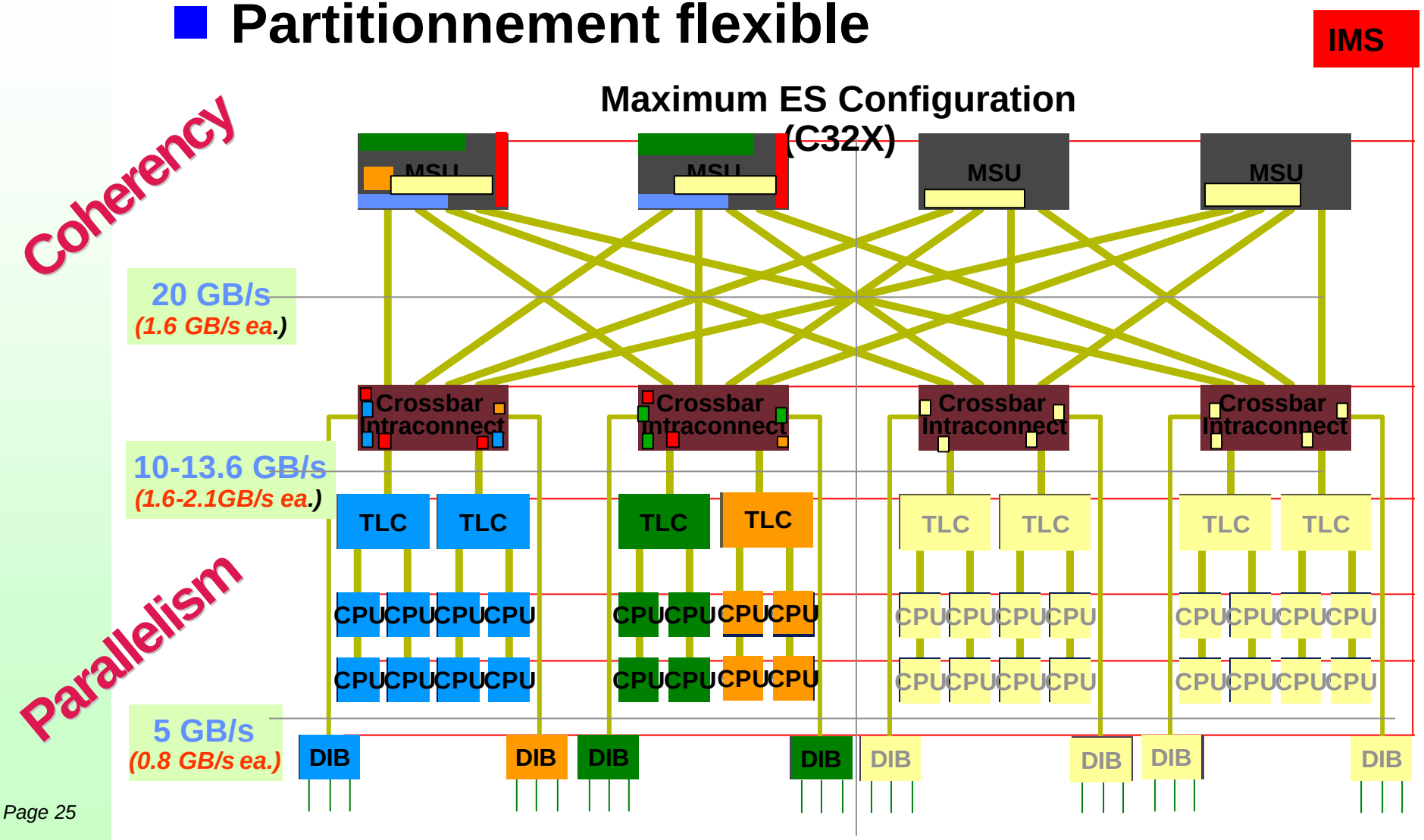
Phased Implementation

# Grands SMP - Unisys(2)

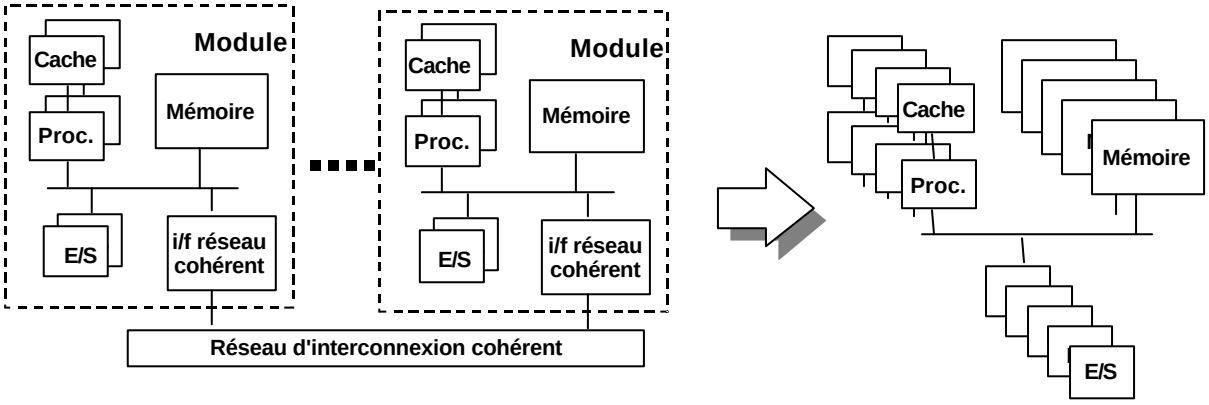


# Grands SMP - Unisys(3)

## Partitionnement flexible



# Principe CC-NUMA




**Configuration physique** **Configuration logique équivalente**

- NUMA Factor : Rapport des temps d'accès mémoire  distante - mémoire locale.
- Plus le NUMA Factor est important, plus le logiciel doit tenir compte des spécificités de l'architecture
- Adaptation nécessaire des logiciels (Systèmes d'exploitation, SGBDs,...) aux caractéristiques de l'architecture CC-NUMA. Il doit tenir compte des propriétés de localité :
  - Affinité processus/module
  - Allocation d'espace mémoire à un processus sur le module sur le lequel le processus réside
  - Rapatriement d'un processus sur le module sur lequel il résidait, .....

# Architecture CC-NUMA(2)

## IBM NUMA-Q (jusqu'à 256 processeurs Pentium)

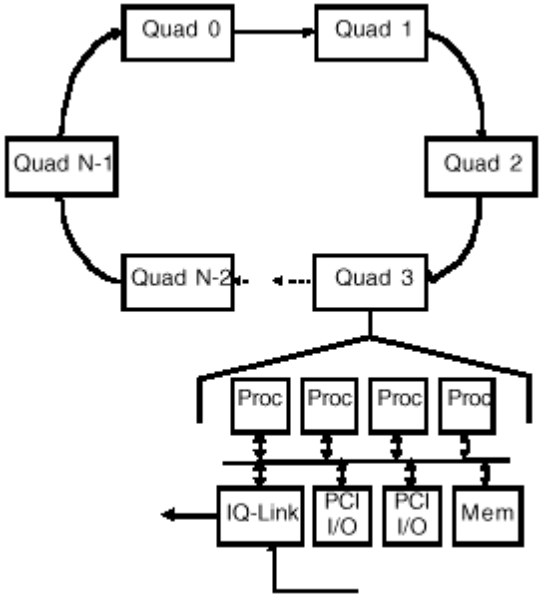



Figure 2.1: NUMA-Q block diagram.

**Module SMP dérivé du SHV d'Intel Pentium III**

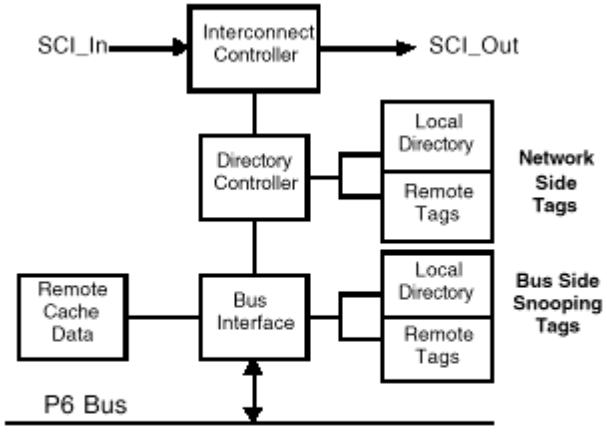



Figure 2.2: IQ-Link board block diagram.

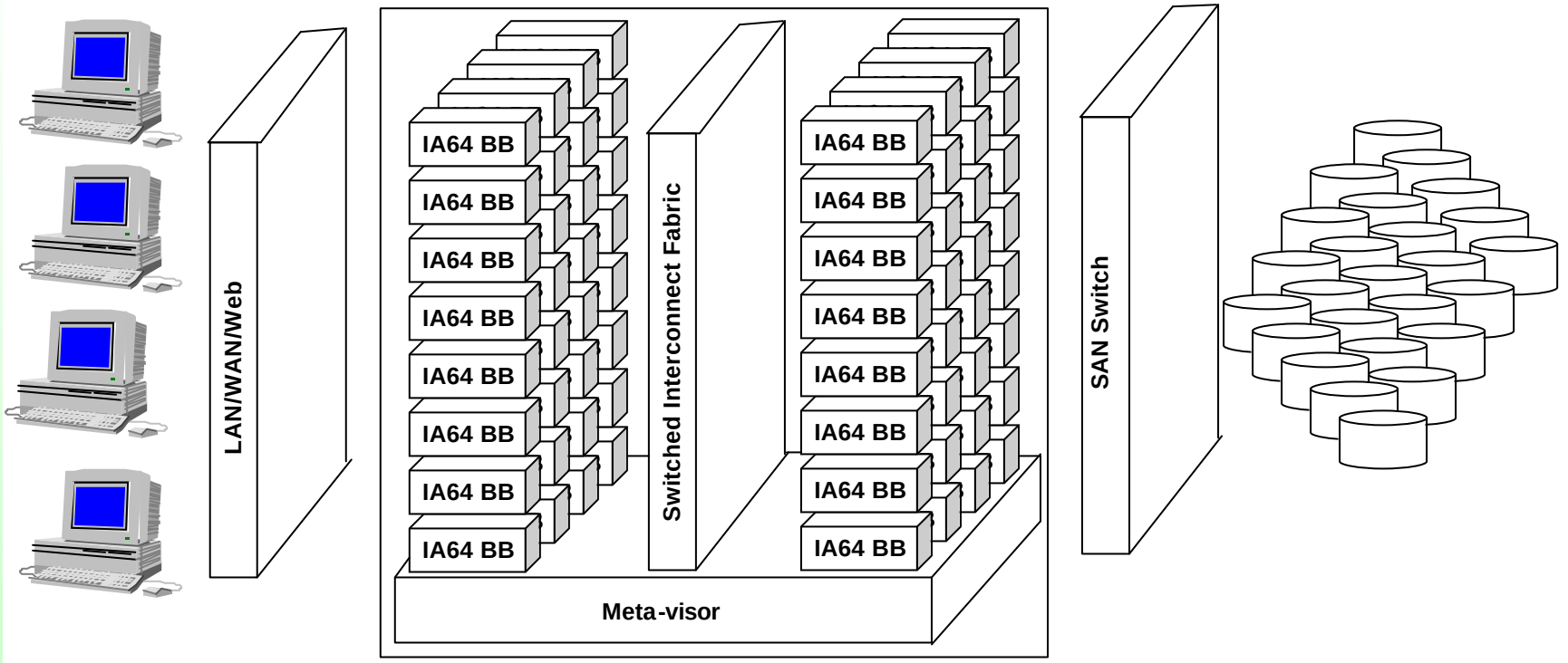
**Architecture du lien inter-modules (IQ-link)**

**Lien fondé sur le standard IEEE SCI (Scalable Coherent Interface)**

**Protocole de cohérence inter-module fondé sur le mécanisme de répertoire (directory-based)**

# Architecture CC-NUMA(3)

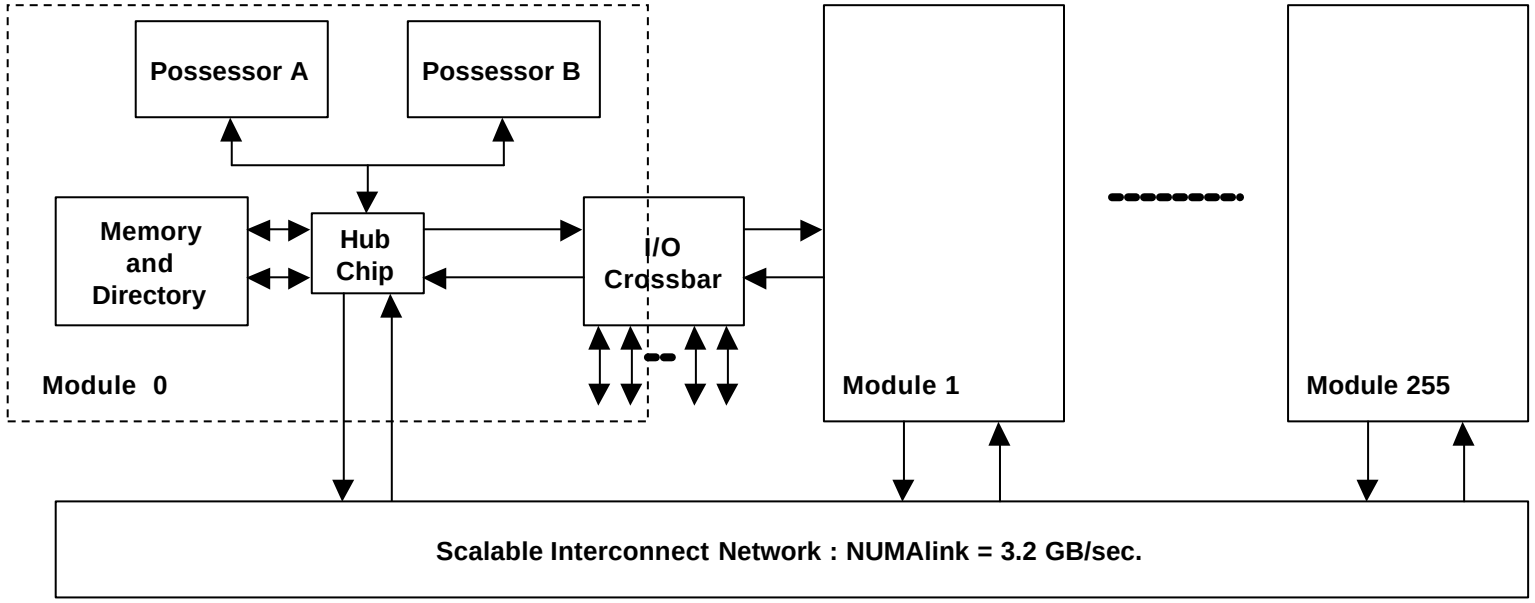
- **IBM NUMA-Q Prochaine génération à base IA-64**




**IA64 BB = IA64 Building Block = quadriprocesseur IA64 + mémoire + E/S**
**256 processeurs IA64 au maximum**

# SGI - Origin 3000 Series

- **Architecture CC-NUMA (appelée NUMAflex) allant jusqu'à 512 processeurs à base d'un module bi-processeur (architecture MIPS)**
- **Architecture du module de base et principe de l'architecture :**




- **Réseau d'interconnexion cohérent - Topologie de type Hypercube (voir ci-après), lien de type 1.6 Go/s bi-directionnel (CrayLink)**
- **Évolution prévue vers IA-64**
- **Migration de IRIX vers Linux (complété progressivement par des caractéristiques de système « haut de gamme » implémentées sur IRIX)**
- **Système pouvant être partitionné**
- **Ajout ou retrait de modules sans interrompre le fonctionnement du système**

# SGI - Origin 3000 Series(2)

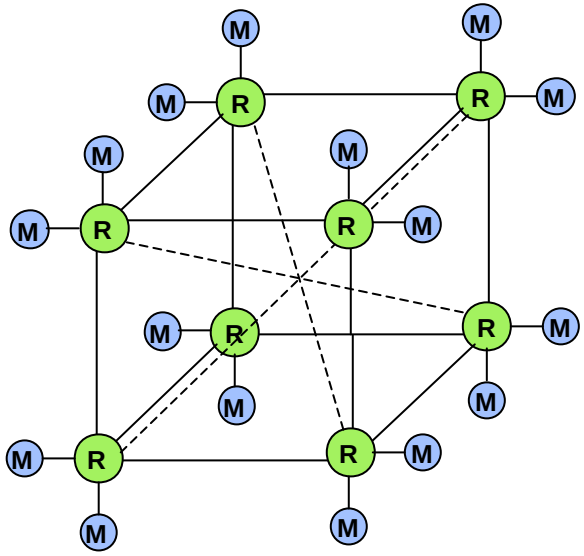
- **Architecture physique modulaire à base de composants (Bricks) :**
  - **C-brick. Module processeur : 4 processeurs et jusqu'à 8 Go. Dans le futur, C-brick fondée sur IA-64**
  - **I-brick. Module d 'entrées-sorties : disque système (FC), 5 cartes PCI « hot plug », CD-ROM, Ethernet (10/100), IEEE 1394, 2 USB, 2 liens Xtown2 (1.2 Go/s bi-directionnel) vers C-bricks**
  - **P-brick. Module PCI additionnel 12 cartes « hot plug ». 2 liens Xtown2 (1.2 Go/s bi-directionnel) vers C-bricks**
  - **X-brick. Module d'extension des entrées-sorties permettant la re-connexion des entrées-sorties des systèmes de la série Origin 2000 (4 cartes XIO)**
  - **R-brick. Crossbar à 8 ports (1.6 Go/s bi-directionnel), 4 ports pour la connexion avec des modules processeur (C-bricks) et les autres pour l'extension du nombre de processeurs**
  - **D-brick. Module disques (12 disques 36 Go puis 18 disques 73 Go) gérés en JBOD (Just a Bunch Of Disks) ou en RAID**
  - **G-Brick. Extension graphique**
  - **Power Bay. Jusqu'à 6 alimentations électriques « hot swap » en configuration N+1**

# SGI - Origin 3000 Series(3)

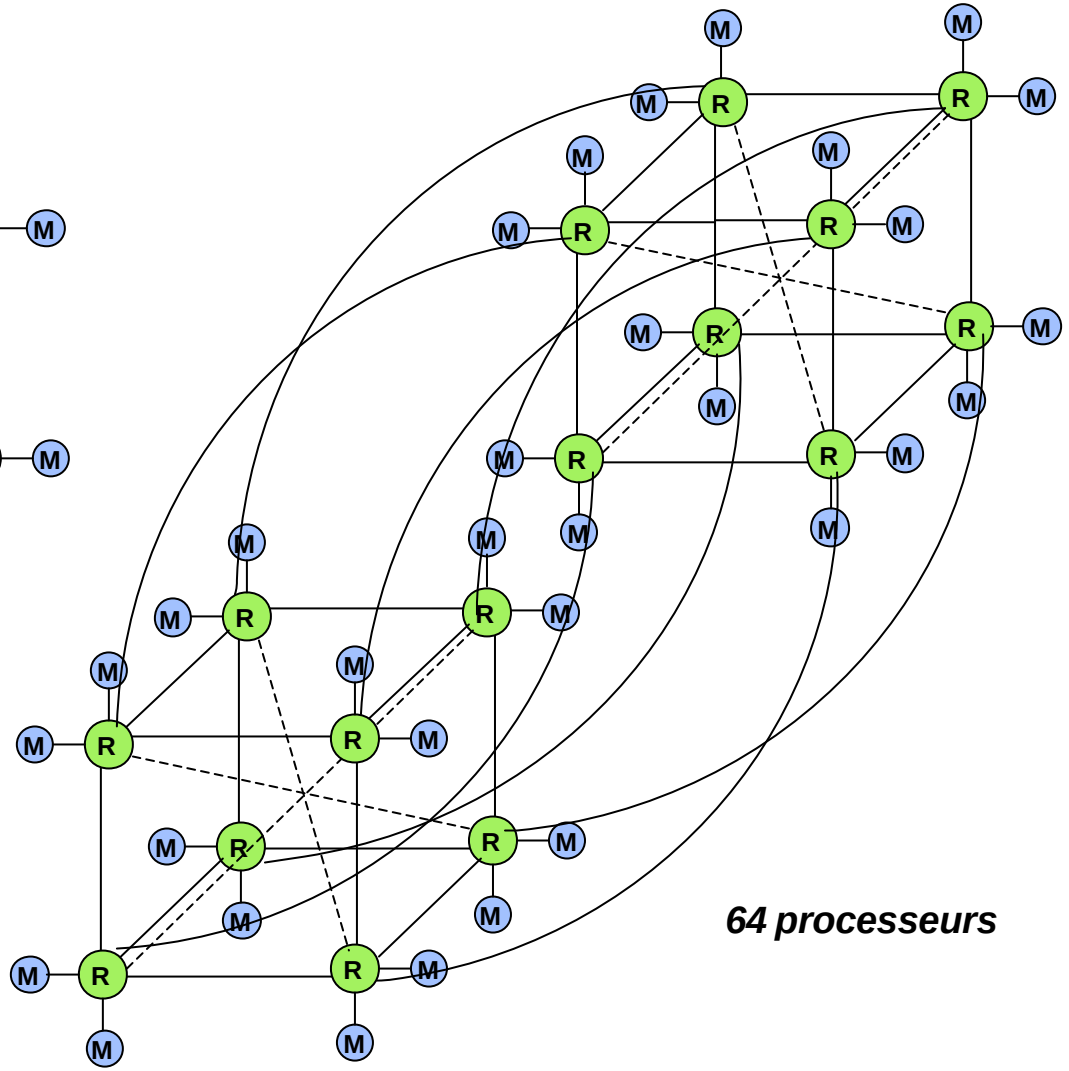
- **Technologie d 'interconnexion :**
  - **2 puces ASIC spécifiques :**
    - **Bedrock : crossbar à 8 entrées et 6 sorties (3.2Go/s) reliant au sein d'un module C-brick : les 2 processeurs, la mémoire et les entrées-sorties;**
    - **Router : crossbar à 6 ou 8 ports utilisé au sein des R-bricks**
  - **Liaison entre les routeurs au moyen de câbles**
- **Topologie de type Hypercube (voir ci-après)**
- **Latences :**
  - **Origin 3000. Un seul chiffre : pour 128 processeurs, le ratio entre le temps d'accès distant et local est de 2.99**
  - **Origin 2000 (génération précédente) :**
    - **Mémoire locale = 310 ns**
    - **2 modules (4 proc.) = 540 ns**
    - **16 modules (32 proc.) = 773 ns**
    - **64 modules (128 proc.) = 945 ns**
- **Protocole de cohérence dérivé de celui du projet DASH**
- **Algorithme de migration de page fondé sur des compteurs maintenus par le matériel**
- **Nouvelle instruction de synchronisation « Fetch and Increment » (fch-inc) pour éviter la contention crée par le couple classique « Load Linked »/ »Store Conditional » (LL/SC).**
  - **Exemple de débit en millions d'opérations par seconde sur Origin 2000 :**
    - **1 processeur fch-inc = 4.0, LL/SC = 6.9**
    - **4 processeurs fch-inc = 6.1, LL/SC = 0.84**
    - **8 processeurs fch-inc = 10.0, LL/SC = 0.23**
    - **16 processeurs fch-inc = 19.3, LL/SC = 0.12**
    - **32 processeurs fch-inc = 23.0, LL/SC = 0.09**

# SGI - Origin 3000 Series(4)

## Topologie d'interconnexion Hypercube




*32 processeurs*

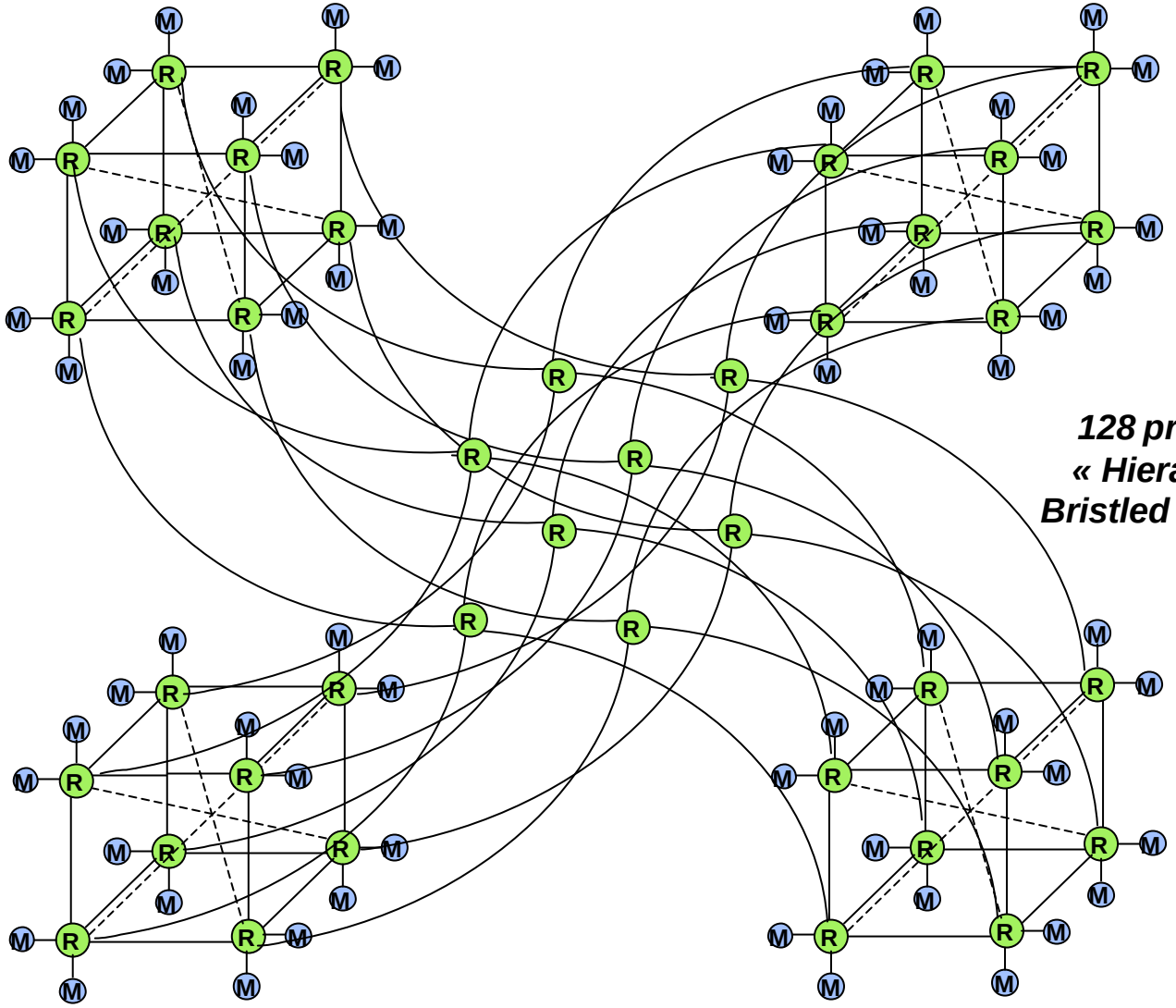



*64 processeurs*

**M = Module bi-processeur**

# SGI - Origin 3000 Series(5)

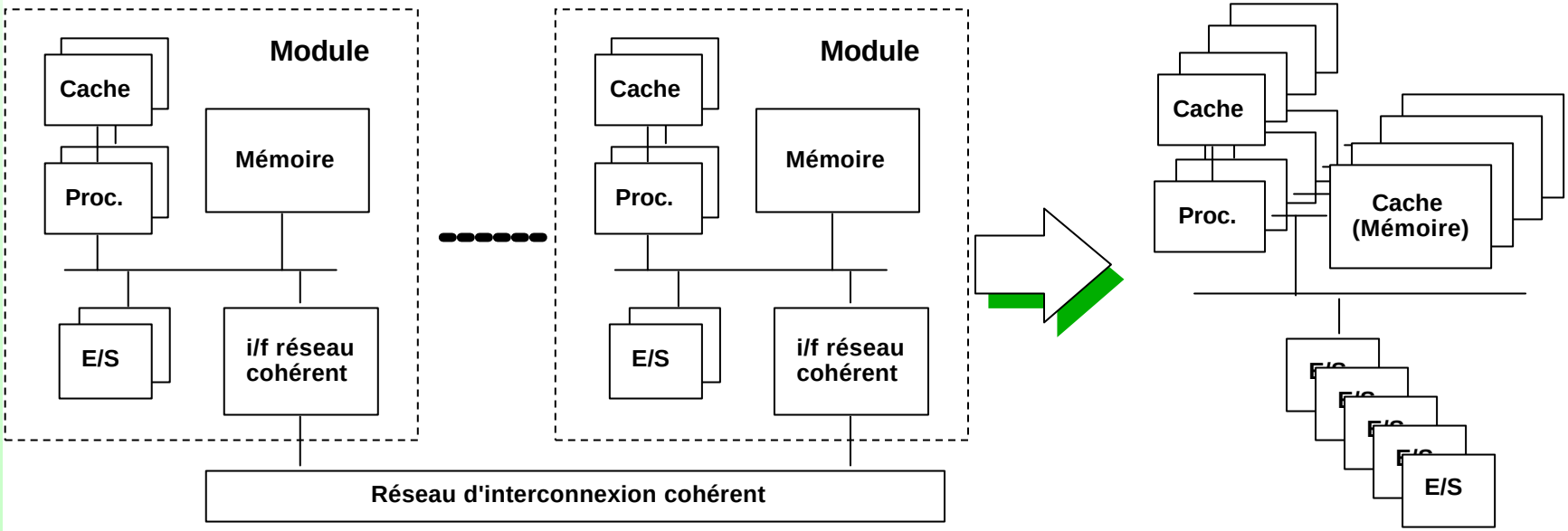
## Topologie d'interconnexion Hypercube (suite)




***128 processeurs « Hierachical Fat Bristled Hypercube »***

# Architecture COMA

- COMA (Cache Only Memory Architecture)
  - **NUMA = Mapping fixe de l'adresse d'une ligne de mémoire sur la mémoire d'un module**
  - **COMA = pas de mapping fixe. L'ensemble de l'espace mémoire est traité comme un cache.**
- Dans l'architecture COMA, la mémoire de chaque module est gérée comme un cache. Les données sont amenées en mémoire locale en fonction des besoins. Un protocole de cohérence est mis en jeu entre les mémoires locales des différents modules.
- Nombreux projets de recherche (FLASH, IACOMA, S3MP, DDM,....). A ce jour, une seule tentative industrielle mais infructueuse = KSR Kendall Square Research.




**Configuration physique** **Configuration logique équivalente**

# Synthèse SMP, CC-NUMA, COMA